Adjusting for Recall Bias in "How Many X's Do You Know?" Surveys

Tyler H. McCormick and Tian Zheng
Department of Statistics, Columbia University, New York, New York

Abstract

Recalling errors project a commonly documented limitation of 'How many X's do you know?' type surveys (see Killworth et al. (2003) for example). These errors often come from respondents' tendency to under-report the number of their acquaintances in larger sub-populations. A calibration curve that corrects for under-recalling is proposed based on the results of Zheng et al. (2006). A description of how to apply this correction to the Zheng et al. (2006) type of model is also discussed.

KEY WORDS: Recalling Bias, Estimating Degree, Social Networks

1. Introduction

Surveys that ask questions of the form "How many X's do you know?" are currently of great interest to social network researchers. Such surveys have been used to estimate the degree[1] and degree distribution of individuals as well as to estimate the size of hard to count populations (Killworth, Johnsen, McCarty Shelley, and Bernard 1998a; Killworth, McCarty, Bernard, Shelley, and Johnsen 1998b).

To this point, surveys of this form have been kept from their full potential because respondents have limited ability to recall accurately their ties with large subpopulations. Consider the question "How many college or university faculty members do you know?" Since the size of this group is rather large for most people in academia, it is difficult to recall each member of the group in the limited time given on a typical survey.

Killworth et al. (2003) documents these effects and proposes several mechanisms to explain under-recall in large subpopulations. One possible explanation is a process that Killworth et al. calls "dredging," whereby a respondent recalls one-by-one the first m acquaintances and then estimates for all groups larger than some size m. This mechanism would, in theory, produce accurate responses for small groups (less than m acquaintances) but less reliable responses for larger groups where respondents are estimating total group size rather than counting specific acquaintances (McCarty et al. 2001). Though this mechanism seems plausible, there is no specific process for determining m or modeling how estimating rather than enumerating would impact the overall accuracy of the results. Additionally, both Killworth et al. and McCarty et al. point out that the relatively short time given to answer each question likely creates difficulty for respondents and is confounded with "dredging."

[1] In social network research, the *degree* refers to the size of an individual's personal network.

In the following sections we propose a calibration curve which corrects for under-recalling and describe how this correction can be applied to models of the type proposed by Zheng et al. (2006). We start by reviewing the development of the Zheng et al. (2006) model. We then show evidence of recalling bias based on the results from this paper and propose a correction. Finally, we describe how to fit this correction to the Zheng et al (2006) model and give a brief data example.

2. Developing the calibration curve

2.1 Recall bias and its effect on model estimates

Zheng et al. (2006) used data collected by McCarty et al. (2001). The data consist of 1,370 adults who were asked to identify the number of acquaintances they have in each of 32 groups including names (e.g. Michael, Christina), occupations and organizations (e.g. commercial pilot, member of the Jaycees), and life events or conditions (e.g. opened a business in the past year, diabetic).

The data were modeled using a multilevel Poisson model with overdispersion. To fit the model, Bayesian computation was carried out under a negative binomial parameterization. More specifically, let y_{ik} be the ith individual's response to the question "how many people do you know in group k." We assume

$$y_{ik} \sim \text{NegBin}(\text{mean} = e^{\alpha_i+\beta_k}, \text{overdispersion} = \omega_k),$$

where e^{α_i} is individual i's degree, e^{β_k} estimates the proportion of ties that link to subpopulation k in the social network and α's, β's and ω's follow upper-level models (The details omitted here. Interested readers are referred to Zheng et al. 2006).

This model has a nonidentifiability since the likelihood depends on α_i and β_k only through their sum. To identify the α's and β's the model is renormalized by adding a constant to all $\alpha_i's$ and subtracting the constant from the $\beta_k's$. One intuitive way of calculating the renormalzing constant is to set

$$\sum e^{\beta_k} = \sum \{\text{population proportion}\}_k. \qquad (1)$$

This is equivalent to assuming that the average degree of individuals in these subpopulations equals the average degree of the population. Obviously, this assumption does not apply to all 32 subpopulations used in McCarty et al. survey. Someone who is a member of the Jaycees, for example, likely has a larger than average degree because of the social nature of the organization. When restricted to the subpopulations defined by the first names, however, this assumption is fairly reasonable.

The above strategy also requires that the acquaintance ties recorded in the survey reflect the distribution of ties in the social network. However, the survey did not accurately measure the social network but rather the *recalled* social network by the respondents. For rare groups, the respondents can recall almost all their ties with these groups. The number of ties to a large subpopulation k is under-recalled. The estimated proportion e^{β_k} from data therefore only estimate the proportion of ties involving subpopulation k in the recalled social network. Consequently,

$$\begin{aligned} \sum e^{\beta_k} &= \sum f\left(\{\text{population proportion}\}_k\right) \\ &\leq \sum \{\text{population proportion}\}_k. \end{aligned}$$

Here, $f(\cdot)$ represents the *recall* function. If the renormalizing constant is computed based on equation (1) and some popular first names, the degrees of the respondents will be underestimated.[2]

Among the first names used in the McCarty et al. surveys, the most popular name is Michael, representing 1.8% of the population. For someone whose personal network size is 600, he is expected to know $600 \times .018 \approx 11$ Michaels. Though imaginable, it is difficult to recall 11 Michaels during the limited amount of time of such a survey; therefore, the actual reported count is likely to be much lower. In fact, in the McCarty et al. data, respondents reported knowing an average of just under 5 Michaels.

In Figure 5 of Zheng et al. (2006), it is shown that for rare names the estimated β_k's and the log population proportions fall closely to the line with slope 1. As the population size increases, the slope of the regression line between β_k and log population proportion is approximately 0.5. Killworth et al. (2003) also discovered this phenomenon. In their explanation, they propose a model for the expected number members of a subpopulation recalled as a fraction of the total number known. They then demonstrate that this model is well approximated by a square-root curve.

To accommodate recall bias in the McCarty et al. data, the normalization in Zheng et al. (2006) is based on the rarest names in the data (such as Jacqueline, Christina, and Nicole) with a correction for the fact that these names are female and that people tend to know more individuals of their own sex.

2.2 Derivation of the calibration curve

The choice to use the rarest names in Zheng et al. (2006) was somewhat arbitrary and was from visual checking based on Figure 5 in Zheng et al. (2006). In this section, we propose to use a calibration curve fitted to all 12 names to adjust for under-recalling as a function of subpopulation size.

Let e^{β_k} be the proportion of ties in the social network that involve individual in subpopulation k. And let $e^{\beta'_k}$ denote the proportion of ties in the *recalled* social network that involve subpopulation k. Assume $\beta'_k = f(\beta_k)$ and $f(\cdot)$ is an increasing function.

[2] In Zheng et al. (2006), we observed that the average degree is about 384 while using all 12 names to normalize and then becomes 739 while using the rarer names.

Based on our observation and also independent discussion by Killworth et al. (2003), we assume that

$$\begin{aligned} f'(x) &\to 1 \quad \text{as } e^x \to 0 \quad (x \to -\infty) \\ &\to \tfrac{1}{2} \quad \text{as } e^x \to 1 \quad (x \to 0). \end{aligned}$$

To simplify the inference, we assume that $f(x) = x$ for small populations with proportion as small as $e^x = e^b$ ($b < 0$) and $f'(x)$ decreases as x increases (at most) to $\frac{1}{2}$ as x goes to zero. More specifically, we assume

$$f'(x) = \frac{1}{2} + \frac{1}{2}e^{-a(x-b)}, a \geq 0, \text{ for } x \geq b,$$

where a controls how fast and how close $f'(x)$ approaches $\frac{1}{2}$.

This gives us

$$f(x) = b + \frac{1}{2}(x-b) + \frac{1}{2a}\left(1 - e^{-a(x-b)}\right).$$

In this paper, we use $b = -7$, which corresponds to subpopulations that are $< .1\%$ of the population and a is to be fitted using β_k originally estimated and the population proportions of first names. This is because, as discussed earlier, we assume that in the absence of recall bias, $\beta_k \approx \{\text{population proportion}\}_k$ on average. Incidentally, we found that an a of approximately one yielded the best fit.

Figure 1 plots the original estimates of β_k against the the known size of the twelve first names used in McCarty et al. data. This is a reproduced version of Figure 5 in Zheng et al. (2006) with the calibration curve. If there were no recall bias, points in Figure 1 should scatter about the y=x line. In Figure 1, however, we see that the y=x line is an acceptable fit for the rarest names, but becomes less reasonable as the size of the subpopulation increases.

3. Modeling using the calibration curve

The calibration curve can be included in the model suggested by Zheng et al. with only minor modifications. We simply adjust β_k in the negative binomial mean by applying the calibration curve.

$$y_{i,k} \sim \text{Neg Bin}(\text{mean} = e^{\alpha_i + \beta'_k}, \text{overdispersion} = \omega_k)$$

where $\beta'_k = f(\beta_k)$.

By including the calibration curve, the estimated parameters remain $\beta'_k s$ and their magnitude has been corrected for recalling.

To see the effects of this correction, consider Figure 2. This figure shows the same log-log plot as in Figure 1, except we have now also included our estimates of β_k after adjusting for recall.

Two things are worth noting on this figure. First, the estimates produced using the recall correction fall generally along a line with a slope of one, indicating that our estimates increase proportionally as the subpopulation size increases. Second, notice that the size of the correction is dependent on the subpopulation size. Smaller subpopulation estimates are corrected less than larger ones, an observation that is consistent with the idea that recall is more accurate in smaller subpopulations.

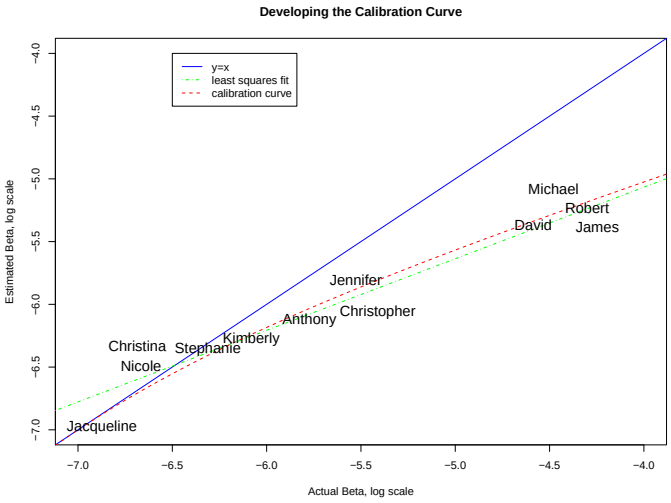

Figure 1: The solid line is the $y = x$ line while the light green line is a least-square regression line fitted that has a slope of 0.53. The red broken line indicate best-fit calibration curve that captures both the $y = x$ pattern at the lower end and the diminishing recall at the higher end of x.

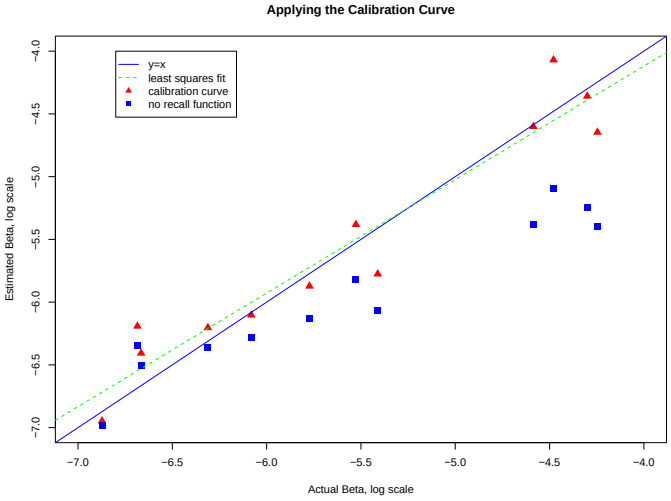

Figure 2: The estimated subpopulation proportions for the twelve names plotted against values obtained from the Census before and after applying the calibration curve (both on the log scale). Blue squares are estimates without the calibration curve and red triangles are recall corrected.

4. Discussion and Conclusion

Though the calibration curve performs well when comparing estimates based on names, researchers are often interested in other categories. It is possible that the mechanism used by individuals recalling names is different than for recalling other categories. If this is the case, the calibration curve may not adjust adequately.

Additionally, Killworth et al. (2003) note that recall bias is only one potential source for inaccurate responses to 'How many X's do you know?' questions. Additional bias is known to come from barrier effects (some respondents are prevented from knowing members of the subpopulation group) and transmission effects (respondent knows someone in a subpopulation but is stopped from knowing that they are in that subpopulation). The calibration curve does not account for these effects, though additional modifications to the Zheng et al. model proposed by Salganik, McCormick, and Zheng can address these issues.

In this paper we have proposed a method for addressing recall issues that are a major limitation of using 'How many X's do you know?' type questions to estimate personal network size. Our calibration curve is derived from observations made in previous independent research and observations made in Zheng et al. (2006). We also incorporate the calibration curve into the Zheng et al. type model and demonstrate that the effectiveness of the curve using the McCarty et al. (2001) data.

The calibration curve we propose in this paper is based on the 12 names in this particular data. A potential extension of this work could include integrating a routine to estimate the optimal value of a to fit the calibration curve seamlessly to different sets of data.

References

[1] Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., and Bernard, H. R. (1998a). A Social Network Approach to Estimating Seroprevalence in the United States. *Social Networks* **20**, 23–50.

[2] Killworth, P. D., McCarty, C., Bernard, H. R., Johnsen, E. C., Domini, J., and Shelley, G. A. (2003). Two interpretations of reports of knowledge of subpopulation sizes. *Social Networks* **25**, 141–160.

[3] Killworth, P. D., McCarty, C., Bernard, H. R., Shelly, G. A., and Johnsen, E. C. (1998b). Estimation of seroprevalence, rape, and homelessness in the U.S. using a social network approach. *Evaluation Review* **22**, 289–308.

[4] McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., and Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization* **60**, 28–39.

[5] Salganik, M. J., McCormick, T. H., Zheng, T. (2007) Efficiently estimating personal network size. *Working Paper, Department of Statistics, Columbia University*

[6] Zheng, T., Salganik, M. J., Gelman, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Assocation* **101**, 409–423.